NCore: Architecture and Implementation of a Flexible, Collaborative Digital Library

Dean B. Krafft, Aaron Birkland, Ellen J. Cramer
Cornell University
June 19, 2008
Presentation Overview

- What is NCore?
- Data Model and Architecture
- Aggregations: Organizing the Library
- Services and Management Tools
- End User Tools
- Status and What’s Next
- Summary
NCore: The Technical Vision

- Provide support for communities across a broad range of disciplines, education levels, and degrees of engagement
- Enable the library as a shared, collaborative, contributory space
- Support the creation and display of context around library resources to enhance discovery, use, and understanding
- Put the library in the path of the user, enabling them to easily and comfortably integrate the library into their normal workflow
- Release NCore as an open-source platform for developing digital library tools
What does NCore provide?

• A digital library architecture that contains
  – References to STEM resources
  – Metadata that describe those resources
  – Ways to organize, interrelate, and annotate resources

• Back-end tools and services to support the creation, organization, and indexing of resource references and metadata in the library

• End-user tools that allow:
  – Discovering, creating, and organizing library resources and metadata
  – Creating context, relationships, and annotations for the materials in the library
NCore in Operation

- NSDL Data Repository contains 6 million digital objects: 3.4 million resources and 2.6 million metadata records
- 150,000 metadata/resource record updates are harvested from over 130 collections each month
- Production services run on 9 Dell RedHat Linux servers, including 3 repository servers with one leader and 2 real-time replicating followers
- Search indexing service regularly crawls and indexes 700,000 STEM resource pages
- Search service handles approximately 5 million queries/month
- Expert Voices blogosphere currently contains 46 blogs and 1524 posts with 749 registered users
The Stacks (Repository)

The NSDL Data Repository (NDR), implemented as a set of digital objects and relationships in a Fedora repository.
NCore specializes Fedora

- Multiple Object Types:
  - Resources (with local or remote content)
  - Metadata
  - Aggregations (collections)
  - Metadata Providers (branding)
  - Agents

- Relationships with arbitrary graph queries:
  - Structural (part of)
  - Annotation (relates to)

- Maintains Fedora’s independence of datastreams - allows other metadata and content to coexist with NCore
Overview: The Core of NCore

• High level object model built on top of Fedora objects
• Middleware that serves several functions:
  - Exposes a high-level API centered around NCore model
  - Serves as a "host environment" for NCore objects in Fedora.
  - Implements business rules, referential integrity, authorization.
Motivation: The NSDL Use Case

- High volume of data from many (hundreds) of independent sources
- Majority of resources and metadata come from automated processes
- In general, NSDL does not have any control or influence on the content of these sources.
- NSDL does wish to present and enrich the multi-sourced context around resources
Motivation: Requirements

• Capability for some editorial leeway beyond wholesale selection/omission of collection.

• Support case where there is no a priori authoritative description of a resource. Often resources originate from one party (who has nothing to do with the NSDL), selected, described, and contributed by several others (who have nothing to do with the creator of the resource, each other, or perhaps even the NSDL).

• Creative ways for user/application select and use resources, or selections of resources.
Motivation: More Requirements

- Allow "the library" to exist as a selected subset of resources in the library
- Allow applications to store and organize data in the repository for their own use, without affecting anybody else

➡ Corollary: Allow for other "libraries" to be constructed within the same repository, incidentally sharing some data, but ultimately operating orthogonally.

In other words, you control your view of the data, but not necessarily the data itself.
Implementation: Key Points

• Separation of data into an atomistic model
  – Metadata and described resource(s) are separate objects
  – Two different "planes" for data to exist in: Graph and relationships between atomistic business objects in NCore, and domain-specific assertions in contributed metadata.

• First-class aggregations, coupled with a simple set of rules and behavior, are an excellent way to realize most requirements.
Implementation: Properties of Aggregators

- May be described by (anybody's) metadata
- May have heterogeneous membership unless otherwise restricted
- Have controlled membership
  - Each aggregation is "owned" by exactly one Agent
  - Permissions to add/remove items are separate from ownership
- Any object may be a member of arbitrarily many aggregations
Implications

• Provenance
• Nested aggregations and structures
• Delegation of Authority & transitive trust
• Views of the repository
• Defining and managing a library through aggregations
• Integrity of selections through rules and authorization (aka: why we don't just consider aggregations a property expressed in metadata)
Implementation: Uses of Aggregators

Ultimately, aggregators are simply used to define groupings of objects, and make assertions about these groupings (sound familiar?)

- Defining a "Collection"
- Identifying the constituents of a complex objects
- Creating persistent lists of resources for use by an application, or for sharing
- Enumerating the resources, metadata, collections, or other aggregations considered "in" (or specifically not in) a "library"
Back-end Services

- NDR API: A REST-based web services interface to the NDR
- Ingest: OAI-PMH metadata aggregator
- OAI-PMH server for library metadata
- NSDL Search Service
- Digital Discovery System
NCore: NDR API

- Uses REST calls for all interactions
- Specializes Fedora for NDR objects/relationships
- Disseminations allow combining metadata from multiple sources, or related content
- Authentication: Requests signed with private key associated with an agent
- Authorization: Agent can become a metadata provider or aggregator; can create resources
OAI-PMH Services

- Harvesting and Ingest
  - Automated process: Harvest trigger files created by scheduling system
  - Full logging with email feedback to provider
  - Automated rescheduling

- Repository OAI-PMH serving
  - Uses Fedora proai service to index Dublin Core datastreams in metadata objects
  - Collections serve as OAI-PMH sets
  - RDF relationships can be expressed and served as metadata
OAI-PMH Automated Harvesting

- Collections validate their OAI-PMH server
- CI registers collection (NCS)
  - harvest schedule, baseURL, set information…
- Full harvest initiated
- Subsequent incremental harvests according to schedule
  - automated emails if problems
NSDL Search Service

- Based on Lucene/Nutch
- Service exposes full power of Lucene queries
- Indexes metadata records incrementally harvested from NDR
- Crawls resources on web, indexing full text of resource
- Scales easily to millions of resources
DDS Search Service

- Digital Discovery System Web Service (DDSWS), developed by DLESE/Digital Learning Sciences as part of DLESE in NCore effort
- REST web service interface
- Supports detailed searching on specialized metadata fields (e.g. education level, georeference)
- Efficient for moderate sized collections (order 10,000 records)
- Documentation at [http://www.dlese.org/dds/services/ddsws1-1/service_specification.jsp](http://www.dlese.org/dds/services/ddsws1-1/service_specification.jsp)
Management/Developer Tools

- NCore Collection System (NCS): full XML-driven metadata cataloging/management (developed as part of DLESE in NCore project)
- Strand Maps Web Service: concept/knowledge mapping and visualization
- Shibboleth Identity Management Clients: user authentication
- Planned: NCore Tool Kit (PHP, Java, and Javascript toolkit for NDR API access)
End-user tools

Blogging with integrated NSDL search, resource linking, and publication, based on WordPress MultiUser

Wiki with integrated NSDL search, resource linking, and publication, based on MediaWiki
NCore Model - Extending Existing Open Source Tools

- Create search service plugin
  - find resources to talk about
  - insert links to resources
- Create data repository interaction plugins
  - add new resources to the library
  - add referenced resources to the library
  - add metadata about resources
- Community sign-on (Shibboleth)
- Add administrative support
- Skins/themes
- Build on RSS
Repository Relationships
Status

- NSDL.org and OAI server/ingest in production since 2002
- NDR/NDR API in production since January 2007; v1.1 released on SourceForge in December 2007
- NDR search service in production since January 2007; SourceForge beta release March 2008
- Expert Voices in production since early 2007; plugins released on SourceForge in March 2008
- NSDL Wiki in production now; plugins release on SourceForge in March 2008
- NCS in production now; released on SourceForge in March 2008
What’s Next in 2008?

• Creation of NCore Toolkit - Java, PHP, and Javascript tools to simplify NCore access and integration

• Integrating Instructional Architect system for creating lesson modules into NCore (joint with Mimi Recker - Utah State)

• Registering RSS feeds to support ingest, recommendations, and bookmarking/folksonomic tagging systems such as del.icio.us and Nature Publishing’s Connotea
Summary - Why NCore?

- It is a platform that can flexibly support a range of digital library and repository applications
- It can layer over existing repositories
- It can scale up to large libraries - certainly tens of millions of resources
- It comes with a significant existing suite of tools and services
- It has been demonstrated to work in a large production library
- It is available in open source from SourceForge under the Educational Community License (ECL 1.0)
- Funding for NCore (as part of NSDL Technical Network Services) now guaranteed through 2012
For more information on NCore:
http://wiki.nsdl.org/index.php/Community:NCore

Collaborative Tools:

OnRamp http://onramp.nsdl.org
Expert Voices http://expertvoices.nsdl.org
NSDL Wiki http://wiki.nsdl.org
Acknowledgements

- NSF EHR/DUE - Lee Zia, Program Officer
- NSDL Core Integration Team
  - UCAR: Kaye Howe, PI and Executive Director
  - Cornell: Dean Krafft, PI
  - Columbia: Kate Wittenberg, PI
- Fedora Development Team
  - Cornell: Sandy Payette & Carl Lagoze
  - Univ. of Virginia: Thornton Staples

This material is based upon work is supported by the National Science Foundation under Grants No. DUE-0733600, 0227648, 424671, and 0227888. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.